WAIS Software Technical Description

Version 2.0

October, 1994

A Wide Area Information Server System

WAIS Inc.'s licensor(s) make no warranties, express or implied, including without limitation the implied warranties of merchantability and fitness for a particular purpose, regarding the software. WAIS Inc.'s licensor(s) do not warrant, guarantee or make any representations regarding the use or the results of the use of the software in terms of its correctness, accuracy, reliability, currentness or otherwise. The entire risk as to the results and performance of the software is assumed by you. The exclusion of implied warranties is not permitted by some states. The above exclusion may not apply to you.

In no event will WAIS Inc's licensor(s), and their directors, officers, employees or agents (collectively WAIS Inc.'s licensor) be liable to you for any consequential, incidental or indirect damages (including damages for loss of business profits, business interruption, loss of business information, and the like) arising out of the use or inability to use the software even if WAIS Inc.'s licensor has been advised of the possibility of such damages, because some states do not allow the exclusion or limitation of liability for consequential or incidental damages, the above limitations may not apply to you. WAIS Inc.'s licensor's liability to you for actual damages from any cause whatsoever, and regardless of the form of the action (whether in contract, tort (including negligence), product liability or otherwise), will be limited to $50.

For more information about WAIS products, contact **info@wais.com**

For WAIS customer support, contact **support@wais.com**

WAIS Inc.

1040 Noel Drive

Menlo Park, California 94025

(415) 617-0444

Printed in the United States of America.

"Millions of people already use the Internet. WAIS is an important tool helping people navigate through the vast oceans of information of the net, and WAIS Inc. is an important pioneer in building the tools which open new information markets."

Mitchell Kapor, chairman of the Electronic Frontier Foundation and founder of Lotus Development Corporation

Table of Contents

Chapter 1 Introduction 1

What is WAIS? 1

The WAIS Architecture 1

What is a WAIS Network Publisher? 2

What You Need to Become a Network Publisher 3

Example Uses of WAIS 4

WAIS Incorporated 5

Technical Description Overview 6

Chapter 2 The WAIS Database 9

What is a WAIS Database? 9

The Parser 9

The Indexer 14

Chapter 3 The WAIS Server 19

Server Operation 19

The Search Engine 19

Query Reporter 23

Security 24

Monitoring and Usage Reports 25

Chapter 4 The WAIS Protocol Suite 27

WAIS and Standards 27

Characteristics of the Protocol Suite 27

Components of the Protocol Suite 28

Chapter 5 The Extendable WAIS Server 31

WAISgate Connects WAIS with Web 31

Serving Multi-Databases 31

Filters Extend Your WAIS Server 32

Custom Parser Toolkit 32

Client Toolkit 33

Chapter 6 The WAIS Forwarder 35

Appendix A Glossary of WAIS Terms 37

Appendix B WAIS References 45

WAIS Articles and Publications 45

WAIS Videos 48

Electronic Services 48

WAIS Freeware Information 48

Z39.50 Information and Publications 49

Internet Information 50

1

Introduction

What is WAIS?

WAIS™ (Wide Area Information Servers™) is a network publishing system designed to help users find information over a computer network by simply asking questions. The questions may be expressed in natural language or use literal phrases, Boolean syntax, or specify field values. The information sources may be local or remote. WAIS software allows users to search for and retrieve documents from information sources all over the world.

As organizations become flatter and more geographically dispersed, the WAIS network publishing system offers an efficient method for accessing information electronically over interconnected local and wide-area networks, thereby greatly reducing printing and distribution time and expenses.

The WAIS Architecture

The WAIS software architecture has four main components: the client, the server, the database, and the protocol, as shown in Figure 1. The WAIS client is a user-interface program that sends search and retrieval requests to local or remote servers. Clients are available for most popular desktop environments. The WAIS server is a program that services client requests. Servers are available on a variety of UNIX platforms. The server generally runs on a machine containing one or more information sources, or WAIS databases. The WAIS protocol is used to connect WAIS clients and servers and is based on the NISO Z39.50 Information Retrieval Service and Protocol Standard.

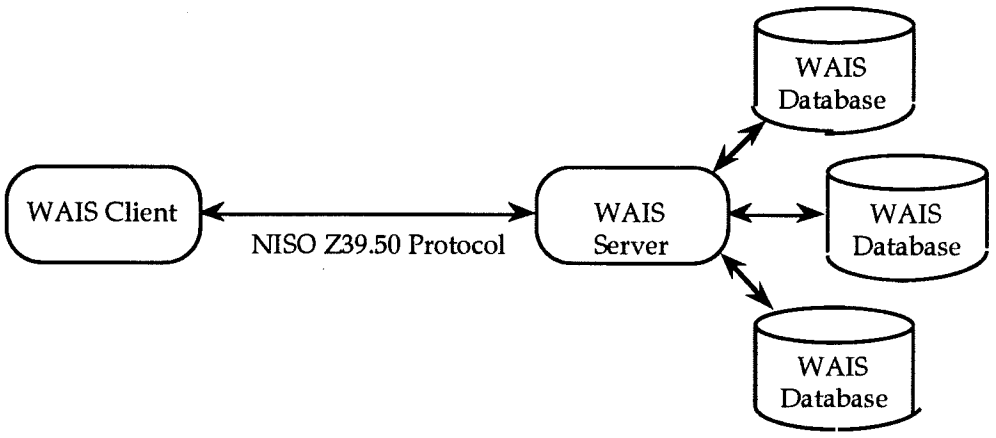

Figure 1: The WAIS Architecture

The goal of the WAIS network publishing system is to create an open architecture of information servers and clients by using a standard computer-to-computer protocol that enables clients to communicate with servers.

The WAIS client-server architecture has many advantages:

Scalability

Its distributed nature allows anyone to set up their own server and become a network publisher. The system can handle thousands of information sources on internets that span the globe, all searchable using standard software.

Efficiency

Current personal computers are high-powered and responsive to the user, and server machines have increased storage capacity and the ability to simultaneously service many users. The client-server architecture lets the client machine interact with the user as a native application on its platform. For example, a WAIS client for Microsoft Windows is a true Windows application and behaves as Windows users expect. Contrast this with most on-line services where a remote server controls what the user sees. The WAISserver, on the other hand, receives its questions in a standard format from all clients and can handle requests without having to recode the response for individual client programs.

Global Communication

The distances involved in global client-server applications often equate to a minimum delay of about one second. Dialup, low-speed lease lines, and wireless connections are typically the most cost-effective means users have to connect to wide-area networks. If information is transmitted on a character-by-character basis over a slow link, the delay between each keystroke could be intolerable. A client-server system can hide much of this delay by packaging up a significant parcel before sending it from the client to the server.

What is a WAIS Network Publisher?

A WAIS network publisher is an information provider that supplies both a WAIS database and a WAIS server, as shown in Figure 2.

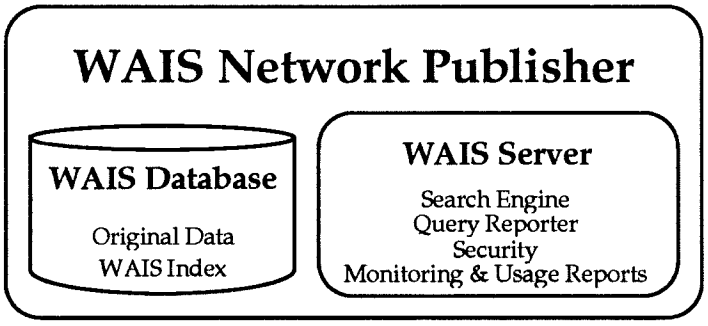

Figure 2: The WAIS Network Publisher

A WAIS database is made up of the publisher's original data collection and a WAIS index to facilitate fast search and retrieval of this data. The WAISserver system is composed of a search engine, a query reporter, a security system, and a monitoring and usage reporting facility. The

WAIS server searches and retrieves documents from the WAIS database. Together, the WAIS database and WAIS server make up a complete network publishing system for information providers.

Because the WAIS architecture is based on a client-server model, the information resides on the server where access can be controlled and usage can be monitored by the network publisher. As a WAIS network publisher, you have control over who has access to your data. Examples for how you might want to control access of your data include:

- Personal use only (e.g., for your personal electronic mail),
- Restricted set of users (e.g., for group or departmental projects),
- Corporate-wide use (e.g., for company resumés and associated photos, company proposals, sales videos and literature), or
- Public use (e.g., for library bibliographies, sales catalogs, marketing literature, public legislation, bulletin boards, and news feeds).

As a network publisher, you can provide your information to each user for a fee, or you can make it freely available. Since each access to your database is monitored by the server, usage patterns are recorded and can be used for statistical analysis or for billing purposes.

What You Need to Become a Network Publisher

It's easy to become a WAIS network publisher. All that is required are the following components:

- A collection of information that you want to publish,
- The WAISserver software,
- A UNIX machine on which the information resides,
- A TCP/IP network to connect clients to the server, and
- An Internet connection (optional).

Each of these components is described in detail below. Keep in mind that most network publishers begin with only a collection of information that they wish to make available to some audience over a computer network. The steps required to become a WAIS network publisher are straightforward, even for those with little or no UNIX experience.

Information

Your information may consist of free-form text, images, and multi-media documents. It does not have to be organized in a structured form, as in traditional database management systems (DBMS). Furthermore, your data collection can be of any size. It may be as small as 1 megabyte, or it may be very large, up to 50 gigabytes in size.

WAISserver

The WAISserver software is used to create an index to facilitate fast search and retrieval of your data. It also supplies the search engine, a query reporting facility, tools for restricting access and monitoring usage, and the WAIS protocol for communicating with clients.

UNIX Platform

The WAISserver runs on most versions of UNIX. WAIS-compatible clients are available on most platforms and operating systems (MS-DOS, MS-Windows, Macintosh, X-

Windows, etc.). Since the WAIS system uses a protocol based on industry standards, servers can communicate with any WAIS-compatible client regardless of the client's platform, operating system, or vendor.

TCP/IP

TCP/IP (Transmission Control Protocol/Internet Protocol) is an industry-standard protocol used to transport information between computers. For a network publisher, it gives the server the appearance of a dedicated connection to a client, and guarantees the integrity of information passed between them. The WAIS protocol is built on top of the TCP/IP protocol, and manages many higher-level functions specific to information publishers and users.

Internet

The Internet is an international computer network used by educational, commercial, military, and government organizations. It is based on the TCP/IP network protocol. To date, the Internet services several million users in dozens of countries, and is growing in geometric proportions. The Internet provides a very wide and broad audience for network publishers. An Internet connection is optional. If you wish to publish your information only to networked clients internal to your organization, you do not need Internet access. Most publishers begin by serving their information internally, and later move to external publication over the Internet.

Example Uses of WAIS

There are hundreds of publicly registered WAIS databases on the Internet. WAIS is used extensively within enterprises and by individuals to provide quick and easy access to information. Here are some examples of how WAIS has been put to use:

Formatted Files

WAIS provides a method for finding and retrieving formatted documents, whether they be phone listings, repair manuals, contracts, status reports, or marketing materials. The text contained in these documents is indexed, while the documents themselves remain unchanged.

Multi-Media

Multi-media documents can be included in a WAIS database by associating a text document with one or more multi-media documents. The text document generally contains textual information about the multi-media documents, and is used in building the WAIS index. When a retrieval is performed, the user has the option of displaying the multi-media documents or the text file associated with them.

DBMS Integration

Some organizations have existing data repositories with processes already in place for collecting, manipulating, backing up and reporting. A WAIS server can be integrated to enable users to quickly retrieve individual records from the DBMS using natural language questions. The actual data remains within the record structure of the DBMS.

Library System

The on-line public access (OPAC) portion of a library system need not be limited to use by only those trained in a proprietary client-server system. Integrating WAIS technology for

public access allows patrons to remotely access MARC records as well as retrieve bibliographic material, citations, or full-text articles and even fully-formatted publications. The WAIS logging and reporting facilities provide a method of integration into circulation control and other library system modules.

Mail Archives

WAIS provides a robust environment with which to search through personal or group electronic mail archives. During a search, WAIS returns the subject line of the mail messages most relevant to the user's question, and then extracts the appropriate mail message when the user performs a retrieval.

WAIS Incorporated

WAIS Incorporated, or WAIS Inc., promotes interoperable WAIS client-server systems. WAIS-compatible systems achieve interoperability by conforming to open national and international standards. As standards change with technological advances, WAIS Inc. maintains continued conformance to these changing standards. WAIS Inc. actively encourages the development and use of WAIS-related products and services by commercial, government, military, and research institutions.

WAIS Inc. offers WAIS-related products and services. This section briefly describes the products and services offered by WAIS Inc.

WAIS Inc. Products

The WAIS Inc. product line provides server software for WAIS-compatible clients. The products include the WAISserver™ and the WAIS Forwarder™:

WAISserver

The WAISserver product is designed for the commercial network publisher whowants to distribute information over wide-area networks.

WAIS Forwarder

The WAIS Forwarder product provides access to WAIS servers from within a secure network. The WAIS Forwarder is appropriate for secure sites connected to an external network, such as the Internet, through a firewall machine.

WAISgate

The WAISgate product provides a gateway for communication between WAIS and the popular World-Wide Web. WAISgate allows Web clients to access WAIS data sets and brings WAIS search capabilities to Web data.

Custom Parser Toolkit

The Custom Parser Toolkit allows WAIS administrators to develop unique parse formats specific to their data, if necessary.

Client Toolkit

The Client Toolkit enables WAIS client writers to get a fast start. It also makes it easy to take advantage of the latest protocol suite, Z39.50-1992.

The WAISserver and WAIS Forwarder products are available on most popular UNIX platforms.

WAIS Inc. Services

WAIS Inc. provides a variety of optional customer services. These include installation and training, product trial evaluations, support and maintenance agreements, and alliance support and development programs.

Installation and Training

WAIS Inc. provides on-site start-up support and system administrator training.

Evaluation Trial

A two-month trial evaluation period is arranged where a WAIS server is set up at the customer's site, or the customer's data is mounted at a WAIS Inc. facility.

Custom Database Integration

WAIS Inc. provides custom integration to a customer's database management system. The custom integration permits a WAIS server to index and retrieve documents directly from the database management system. Custom parser development is also available. WAIS Inc. provides a free evaluation of a customer's database integration needs, and can recommend an integration strategy.

Support and Maintenance Agreement

Customers receive updates, bug fixes, technical support, and training information from WAIS Inc. on an on-going basis.

Alliance Support and Development Program

WAIS Inc. or their licensed representatives provide dedicated personnel for on-going administration, support, development and customization of WAIS Inc. servers.

Overview

The remainder of the *WAIS Software Technical Description* presents the technical details of the WAISserver product from WAIS Inc. The sections include:

Chapter 2 The WAIS Database

Defines what a WAIS database is, and how the parser and the indexer programs are used to create a WAIS database. It describes the input file formats of the parser, the parser's output, and the steps required to create a new parser. It also details the components of the indexer, how an index is used in a search operation, and many additional features of the indexer.

Chapter 3 The WAIS Server:

Describes the operation of the server, the search engine, the query reporter, the security system, and the monitoring and usage report facilities. It also describes the server extension capability using external filters.

Chapter 4 The WAIS Protocol Suite

Describes the WAIS protocol suite as based on both nationally and internationally-accepted standards. It also explains the characteristics and components of the WAIS protocol.

Chapter 5 The WAIS Forwarder

Explains the operation and features of the WAIS Forwarder product.

Appendix A Glossary of WAIS Terms

Contains a dictionary of WAIS-related terms and definitions.

Appendix B WAIS References

Includes a listing of WAIS Inc. literature, WAIS-related articles, publications and videotapes, electronic services, and WAIS freeware information, Z39.50 publications, and Internet information.

2

The WAIS Database

What is a WAIS Database?

A WAIS database is made up of two components: a collection of information, and a WAIS index of this collection. The collection of information is generally referred to as the original data, or just data. It is supplied by you, the network publisher. The WAIS index is a set of files generated by the WAIS programs. It facilitates fast search and retrieval of the information stored in the database. Taken together, the original data and the WAIS index are the essential ingredients making up a complete WAIS database system.

The original data is made up of a set of documents, headlines, and words. A document is the smallest retrievable element of the data collection. For example, a data collection may contain a volume of journals, where each article of the journal is a separate document. Another example is electronic mail, where each mail message is a different document. Each document may be represented in several formats: text, image, or video, or a combination thereof. In a data collection of commercial product catalogs, for instance, each document may contain a textual description of the product, a picture, and possibly a short sales video.

Each document contains a headline. The headline is used to convey the main idea behind the document. When a client sends a question to a server, the server responds with a list of headlines that represent the most relevant documents related to the client's question. Generally, each headline is automatically extracted for you from your original data collection.

In addition to a headline, each document has associated words. These words constitute the text of the document. Together, the headline and the words are used to determine how relevant a document is to a client's question.

A WAIS database is constructed by using the WAIS parser and WAIS indexer programs to create a WAIS index. The parser reads the original data and separates it into documents, headlines, and words. The indexer takes the information generated by the parser and creates the WAIS index.

The Parser

The WAIS parser is a program that separates a collection of documents into components consisting of a headline, field information, and words. The WAIS parser is actually made up of a large number of parsers for handling many different kinds of document formats.

The WAIS parser was intentionally designed to be a program distinct from the WAIS indexer program. Separating the WAIS parser from the WAIS indexer creates a more modular and maintainable environment for incrementally developing new parsers without any need to modify the WAIS indexer. Adding a new parser is as easy as defining a small number of new functions.

This section details the input file formats for the parser, the format of the WAIS parser output, and the steps required to add a new parser.

Parser Input

The input to the WAIS parser is a set of documents making up your original data collection. The data can consist of collections of text, audio, and visual documents in many different file formats. Additionally, the data may be structured in a wide variety of ways. For example, the data may be organized in a single file, across multiple files, or as multiple files in different directory trees. The WAIS parser reformats the data and outputs a single common format acceptable to the WAIS indexer.

The WAIS parser needs to know three items: what documents need to be parsed, how the parser should read these documents, and how the client should display these documents. Each of these is described in more detail below.

Document Specification

The documents in your data collection are specified by filename. If multiple files are specified, the WAIS parser parses each file, one at a time, in the order they are listed. If a directory is specified, all files in that directory are parsed in alphabetical order by filename. With a special option to the parser, the parser can also recursively traverse each directory tree in a depth-first manner.

Parse Formats

When the WAIS parser encounters a new file, it must know how to read the information contained in the file. For example, it needs to know if the file contains a single document, or a set of documents. The parser must also be able to distinguish the headline, the field information, and the words of the text. Generally, most files use one of the following parse formats supported by WAIS Inc:

dash

The **dash** parse format is useful if a single file contains multiple distinct documents. In the dash format, each document is separated by a row containing a minimum of 20 dash characters, "-". The line following the dashed line is expected to contain a headline, followed by the text of the document.

dvi

The **dvi** parse format is for Device Independent Printer Output files. The filename is used for the headline, and the contents of the file supply the words of the document.

filename

The **filename** parse format treats each file as a single document, and uses the filename as the headline. The contents of the file however are generally not parsed. This format is useful for data collections made up of many individual binary files, for example, whose contents are not words.

first-line

The **first-line** parse format specifies that each file contains a single document, and that the first non-blank line of the file is the headline, and the remainder of the file is parsed as words.

first-words

The **first-words** parse format is similar to first-line, except that the headline is the first 100 non-whitespace characters in the file.

gif

The **gif** parse format is for CompuServe's popular Graphics Interchange Format files. The file is considered to be a single document where the filename is used as the headline, and there are no words in the document. Image files, such as gif, can be associated with a descriptive text file and considered as a single document by the WAIS parser, indexer, and server.

html

The html parse fomat is for files using the Hyper-Text Markup Language. This is the standard format for files used by World-Wide Web servers and clients.

mail-digest

The **mail-digest** parse format is for standard Internet mail digest files. A mail-digest file contains one or more electronic mail messages, in which each mail message is parsed as a separate document. The subject line of the mail message is the headline, and the body of the message contains the words of the document.

mail-or-rmail

The **mail-or-rmail** parse format is used for UNIX mail files. A mail file is a single file containing one or more electronic mail messages, where each mail message is parsed as a separate document. The subject line of the mail message is the headline, and the body of the message contains the words of the document. In addition, the sender, the receiver, and the date are recognized as field information.

netnews

The **netnews** parse format is used for Internet Network News, where each Network News or Read News file contains one or more news messages. Each news message is parsed as a document, where the subject line is the headline, and the body of the message contains the words of the document.

one-line

The **one-line** parse format is a simple format that treats each line of a file as a separate document. The line also forms the headline for that document.

paragraph

Like the dash format, the **paragraph** parse format is useful if a single file contains multiple distinct documents, or paragraphs. In the paragraph format, each paragraph is separated by one or more blank lines. The first line of each paragraph is the headline which is followed by the text of the document.

pict

The **pict** parse format is for Apple PICT image files. The file is considered to be a single document where the filename is used as the headline, and there are no words in the document. Image files, such as pict, can be associated with a text file and considered as a single document by the WAIS parser, indexer, and server.

ps

The **ps** parse format is for PostScript files. The filename is used for the headline, and the contents of the file supply the words of the document.

source

The **source** file format is a file format generated by the WAIS indexer for the Directory of Servers. The file typically contains information about the database, and is parsed exactly like the text file format.

text

In **text** format, each file is treated as a single document, the filename is used as the headline, and the contents of the file are parsed as words. This format is useful for data collections made up of many individual files. This is the default parse format.

tiff

The **tiff** parse format is for tagged interchange file formats. The file is considered to be a single document where the filename is used as the headline, and there are no words in the document. Image files, such tiff, can be associated with a descriptive text file and considered as a single document by the WAIS parser, indexer, and server.

Display Formats

The display format determines how a client should display a retrieved document. In many cases, the document is a simple text file, and thus has no special display needs. In other cases, the document may be in a specific format that should be displayed with a special display program. For example, suppose a document was generated using Microsoft Word. The client that retrieves this document must be told that the document is in Microsoft Word format in order to display it correctly. To do this, a display format of **MS-WORD** is given to the parser.

The parser associates a display format with each document. The parser passes the display format through to the indexer, which in turn stores it for later use by the server. When a client requests a search, the server responds by sending the client a list of matching documents. For each document, the server sends back a headline, document identifier, and a list of available display types for that document. It is up to the client to decide whether or not it can display this format, and to specify its preferred display type when it retrieves the document.

Examples of typical display formats supported on many existing WAIS clients are listed in Table 1.

Display Format	Description of Format
DVI	Device-Independent Printer Output
GIF	Graphics Interchange Format CompuServe Images
HTML	Hyper-Text Markup Language
MIME	AT&T Multimedia Document
MS-EXCEL	Microsoft Excel Spreadsheet
MS-POWERPOINT	Microsoft PowerPoint Slides
MS-WORD	Microsoft Word Document
PERSUASION	Aldus Persuasion
PICT	Apple PICT Image
PS	PostScript
QUICKTIME	Apple Quicktime Movie
TEXT	ASCII Text
TEXT-FTP	Special FTP File Format
TIFF	Tagged Interchange File Format A Universal Raster Image Format
WQST	WAIS Question Format
WSRC	WAIS Source Format

Table 1: Display Formats

The exact name of the display format you select depends on the display format supported by the WAIS clients that will be accessing your database. For this reason, you may want to check your client software to determine what display formats it supports.

Parser Output

The output of the WAIS parser is a formatted stream that can be piped directly into the WAIS indexer. The basic entity output by the parser is the document. For each document, the parser outputs specific document information, the headline, file information, field information, a date, and the words of the document.

Document Information

For each document, the parser outputs document information containing the location of the document relative to other documents in the data collection, and the names of the file or files that make up the document.

The location of each document is recorded by the parser as follows. Each document encountered by the parser is assigned a unique document identification number, or doc-num. The parser also records the doc-num of the containing or parent document. These identifying numbers are used to determine if the document is a piece of some larger document, such as a section of an article, or a chapter in a book. If the doc-num of a document is the same as the doc-num of its parent document, the document is a stand-alone document. Additionally, the parser records whether or not a document is in an ordered series of documents. Taken together, the doc-num, the parent doc-num, and the sequential document ordering provide the capability for systematically browsing through documents in a database.

As part of the document information, the parser also extracts and records the name of the file or files that make up the document. There may be one primary file containing the text of the document and several secondary files that contain audio or visual information.

Headline

Each document contains a headline, a string of one or more words specifying the main theme of the document. The headline is used for subsequent indexing, searching, and retrieval. The parse format determines how the parser should extract headlines. For example, if the parse format is specified to be "first-line", the parser extracts the headline from the non-blank line of the file.

File Information

Each document is extracted from one or more files. For each file associated with a document, the parser records information about the file. This information includes the date the file was created, the date the file was last modified, the display format of the file, and the location in the file where the document begins.

Date

The parser extracts a date for each document in the data collection. A date consists of a year, month, day, hour, minute, and second component. If a document is made up of several files, the parser records the date of the most important file.

Field Information

A field is a subsection document. For example, a document could have an "author'" field whose value is the name of the document's author. In WAIS, fields allow a user to restrict a search to a subsection of a document. A user may wish to restrict a search to only those electronic mail documents whose subject line is "Generic Record Syntax" and whose sender is "John", for instance.

A WAIS database may contain up to 254 fields. Each field may be identified by several names. In electronic mail documents, for example, the "sender" and the "from" fields are different names for the same field. The parser must be able to recognize the start of each field, and parse its associated value. The parser then sends this information to the output stream for further processing by the indexer.

Words

If a document consists of text, each word in the text is extracted by the parser. The parser outputs each word, the weight of the word, and the location of the word in the file.

The Indexer

The WAIS indexer takes the information generated by the WAIS parser, and formats it for efficient search. The WAIS index's overhead is small, typically one-third to one-half the size of the original data.

This section describes the various components of the WAIS index and explains how the server uses the index during a search operation. It also covers some of the additional features of the indexer, including incremental indexing, customizable stopwords, and stemming.

Components of a WAIS Index

The WAIS indexer creates a WAIS index, which is made up of the following components:

Dictionary

The dictionary contains a sorted list of all the words used in the data collection. Each word points to a corresponding entry in the inverted file.

Inverted File

For each word listed in the dictionary, the inverted file lists all the documents containing that word. Each document listed contains a pointer to a corresponding entry in the document table.

Document Table

The document table contains a record of each document in the data collection.

Headline Table

The headline table contains the headlines of all the documents in the data collection.

Filename Table

The filename table contains a list of the filenames in the data collection.

These files are automatically generated by the WAIS indexer and are sequentially referenced by the WAIS server during a search request. For efficiency, they are stored in binary format. In addition, there are three human-readable auxiliary files used by the WAIS server:

Catalog

The catalog contains a human readable list of headlines and document identifiers for some or all of the documents in the database. This list may be returned to a user whose search has gone poorly, as an aid to help them understand the contents of the database. The catalog is automatically created by the WAIS indexer.

Source Description

The source description describes a WAIS database and its server. It is used by the client to contact the server and search its database. Typical contents are the machine name, IP address, port of the server, the name of the database, and a short description of the database. The source description file is created by the WAIS indexer the first time the data collection is indexed, and is updated by the database administrator.

Access List

The access list contains the addresses of all machines which are allowed to search the database. It is an editable file created by the database administrator to control access to the database.

How the Index is Used

The interaction between the WAIS index files is illustrated in Figure 3. A client process uses information from the source description file to find and contact the server. The server checks the access list to make sure the client has permission to access this database. If so, the server process takes the words from the client's query and looks them up in the database's dictionary file. The dictionary file provides pointers into the inverted file, where, for each word, there is a list of

pointers into the document table corresponding to the documents that contain that word. The information from the inverted file is used to look in the document table, which gives a pointer to the headline table, which in turn gives a pointer to the filename table. Finally, the information from the filename table is used to access the original data. A list of headlines and relevance scores is returned to the client process for display to the end user.

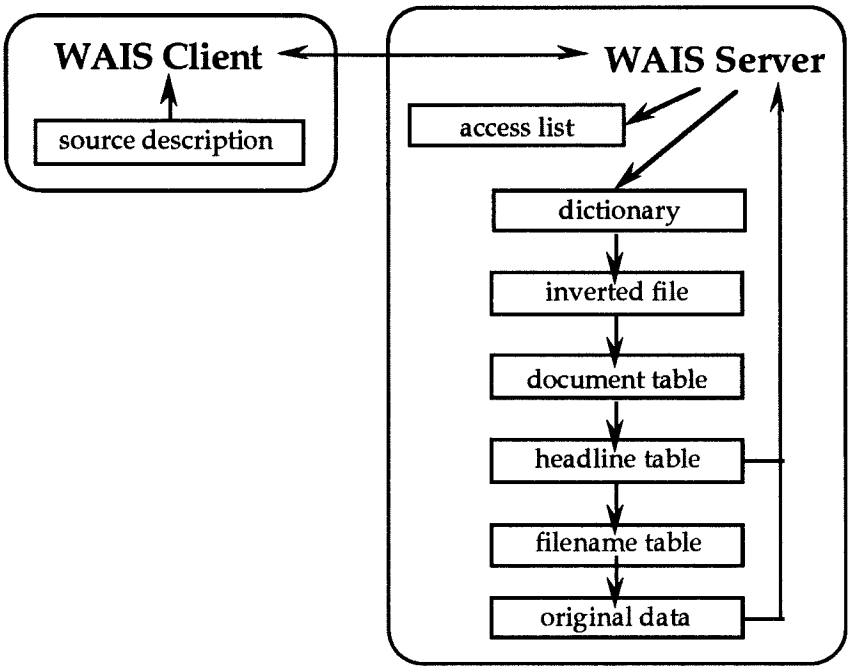

Figure 3: How the WAIS index is used during a search.

Incremental Indexing

The WAIS indexer offers incremental indexing. Incremental indexing allows you to add, modify, or delete WAIS database documents without reindexing the entire database and without suspending user service. Incremental indexing modifies the WAIS index to reflect data changes since the last time the data was indexed. This capability is especially important for network publishers whose data changes often, and whose database size is large.

Customizable Stopwords

A stopword is a frequently used word that, when encountered in a user question, is ignored. For example, since the word "the" commonly appears throughout the English language, it does not help distinguish between documents. Thus it is typically regarded as a stopword. The WAIS indexer includes a list of approximately 300 standard stopwords which can be specially customized for each WAIS database.

Stemming

Stemming is a technique used to automatically derive variations of a queried word. These variations are then used as part of the search. If stemming is used, then when a data set is indexed, word stems are indexed where possible. For example, "dancing," "danced," and "dancer"

would all be indexed as "dance." A question containing the word "dancer", would then turn up documents that may also include "dance", "danced", and "dancing". Two types of stemming are supported: Plural and Porter stemming. Plural stemming attempts to determine the singular form of a word. Porter stemming[2] attempts to find the real base, or stem, of a word and derive any possible alternate variations. The stemming algorithm is selected by the database administrator prior to indexing the database.

[2] Porter, M.F., "An Algorithm For Suffix Stripping," Program 14 (3), July 1980, pp. 130-137.

3

The WAIS Server

Server Operation

The WAIS server is a process that runs on a machine containing a WAIS database, and services requests posed by a client process. The server receives two types of requests from a client: a request to search and a request to retrieve. A search request invokes the WAIS search engine which returns a list of headlines matching the user's question. A retrieve request returns the requested document to the client process.

The operation of the WAIS server is very simple. First a WAIS client initiates a request to a server machine on which the WAIS database resides. Next, the server machine responds by creating a new server process to handle the client's request. And finally, when the client has completed all requests from the database, it closes the connection and the server process terminates. The server also offers a query reporting facility, security, monitoring and usage reporting, and extensions using external filters.

The Search Engine

The WAIS search engine is at the heart of the WAISserver and Workstation products. The WAIS search engine receives a user's question, searches its database for documents most relevant to the question, and returns a relevance-ranked list of documents back to the user. A question is regarded as an expression containing any combination of natural language, quoted literals, Boolean terms, and relevant documents. Other key features of the WAIS search engine include relevance ranking, fielded searches, and right truncation (wildcard matching).

Natural Language

The server can be queried using natural language questions. The server does not understand the question, rather it takes the words and phrases in the question and finds documents that have those words and phrases in them.

```
Tell me about portable computers.
```

is an example of a natural language question.

Literal Phrases

A similar but more specific kind of query asks to find documents that contain one or more exact phrases simply by enclosing them in within quotation marks. This is known as a *literal.* For example, the query

```
"suffering ignorance"
```

returns only documents that contain this phrase. The WAIS search engine treats quoted literals exactly as if each pair of words had the Boolean ADJ operator between them.

Relevance Feedback

Relevance feedback is the ability to select a document or a portion of a document and find a set of documents related to it. For example, suppose you perform a search on a news database with the natural language question "What's going on in personal computers?". Scanning the headlines returned, you see the headline, "Personal Computers in K-12", where you are interested in finding more articles related to this. You can then perform the search again using your original question, selecting this article, or a portion of the article, for relevance feedback. The search engine then returns a new list of headlines for related articles.

In essence, relevance feedback adds more words to the original question. The server analyzes the resubmitted document and determines which of the words in the document are significant in the database; In other words, which of the document's words are useful discriminators setting this document apart from the rest of the documents in the database. It then uses those words to find other documents which share them. The significance of the document's words is less than the original question's words since the user explicitly asked for those words.

One of the primary uses of relevance feedback is to help users quickly focus their search without the need to learn complex query languages. For example, you can use natural language to return a starting list of document headlines, and then use relevance feedback to focus your search on the documents most relevant. Since a WAIS search is fast, you can interactively and iteratively refine your search using a combination of natural language questions and relevance feedback.

Boolean Operators

The Boolean operators, AND, OR, NOT, and ADJ aid in establishing logical relationships between concepts expressed in natural language. These operators are especially useful in narrowing down the search.

AND, &&

> The AND operator is helpful in restricting a search when a particular pair of terms is known. For instance, when searching for documents on the weather in Boston, a question such as "weather AND Boston" would return only those documents that contain both the word "weather" and the word "Boston". Note that the C-like double ampersand operator may be used instead of spelling out the word AND.

OR, ||

> The OR operator is often used to join two different phrases of a Boolean search. A question such as "hurricane OR tornado" would search for all documents containing either the word "hurricane", or the word "tornado", or both. A natural language question is much like having an implicit OR between the words, except that natural language does more work to determine the relevance of words and their relationships in phrases. Note that the C-like double vertical bar operator may be used instead of spelling out the word OR.

NOT

> The NOT operator is used to reject any documents that contain certain words. The question "basketball NOT college" would find all documents containing the word "basketball", that also do not contain the word "college".

ADJ

The adjacent operator, ADJ, is used to ensure that one word is followed by another in the returned document, with no other words in between. For example, "cordless ADJ telephone" returns only documents with exactly "cordless telephone" and not any documents that only contain the words "cordless" and "telephone" separately.

Mixed Natural Language And Boolean Operators

Unique to the WAIS Inc. server is the ability for users to combine natural language and Boolean operators to better target their searches. For example, suppose you were looking for documents about portable laptop computers that are not made by Apple. The question could then be "Tell me about portable laptop computers NOT Apple.".

Relevance Ranking

Each document is scored based on its relevance to a user's question, where the most relevant document has the highest score, or rank. A document receives a higher score if the words in the question are in the headline, or if the words appear many times, or if phrases occur as in the question. A document's score is derived using techniques such as word weighting, term weighting, proximity relationships, and word density. Note that questions made up of natural language, relevant documents, and Boolean expressions are all weighted using these techniques.

Word Weight

If a word in a document is found to match a word in the user's question, the word is assigned a weight, and this weight additively contributes to the overall score of the document. The exact weight that a word receives depends on the emphasis given to the word by the author, and on where in the document the word was found. For example, a word is weighted highest if it appears in the headline, less if the word appears in all capital letters or if the first letter of the word is capitalized, and finally, a word has the least weight if it appears only in the text. The WAIS parser determines word weights as it reads through the original data collection.

Term Weight

Each word used in data collection is assigned a numerical value, called the term weight, based on the frequency of occurrence of that word over all documents in the data collection. Words that occur frequently are not weighted as highly as those that appear less frequently. Very common words are either ignored or diminished in the scoring. For example, since the term, "animal", may occur frequently in many of the documents in a data collection, its term weight is small compared to a term such as "hippopotamus", which may occur only a few times.

Proximity Relationships

Proximity relationships designate that if the words in a natural language question are located close together in a document, they are given a higher weight than those found further apart. The idea behind a proximity relationship is that words found in close proximity to each other in a document more likely contain the same content as that specified in the user's question.

Word Density

The ratio of the number of times a queried word appears in a document to the size of the document is called the word density. It is a measure of how important a queried word is to the overall content of the document. A higher word density results in a higher relevance ranking.

Fielded Search

For data collections whose documents are structured in a semi-regular format, the regular portions of the documents can be tagged by the WAIS parser as fields. Multiple fields may be defined for a single line in a file. Once fields have been tagged, a query sent by a client process can then ask the WAIS server to search for only those documents containing a user-specified value of a particular field. Performing a restricted search based on the value of a field or set of fields is called *fielded search*. In its simplest form, a query containing a fielded search is specified as follows:

field-name = field-value

A query of this form tells the WAIS server to only search for documents whose field, specified by *field-name*, contains the value specified by *field-value*. The WAIS parser can support up to 254 unique fields.

The **mail-or-rmail** parse format is an example of a parse format in which fields are tagged. For this parse format, the WAIS parser detects the "to" and "cc" fields, the "from" and "sender" fields, the "subject" field, and the "date" field.

An example of a question using natural language, a Boolean operator, and fielded search is

company picnic AND from=barbara

In response, the WAIS server would find documents containing messages about a company picnic and a "from" field containing "barbara".

For the date field, a range of dates may be specified, using the syntax

date *comparison-operator value*

The *comparison-operator* may be one of the following symbols: **>**, **<**, **>=**, **<=**, **=**, which mean "greater than," "less than," "greater than or equal to," "less than or equal to," and "equal to," respecively. The *value* argument in a date range query may be a date specification in one of the following formats:

7/4/94	**07/04/94**	**7.4.94**	**07-04-94**
7-4-94	**7-4-1994**	**yesterday**	**today**

As an example of a date range search, the query

date >= 10/14/94 AND name = Tiernan

searches for all documents in which the date field value is greater than or equal to October 14, 1994 and the name field value is "Tiernan".

If the *comparison-operator* is **=**, then the range may be specified using the **TO** operator, as in

date = 1/6/94 TO 6/6/94

here, both ends of the date range are inclusively specified.

Right Truncation

A user can specify right truncation by ending a word with the asterisk ('*') wild card character. This tells the search engine to search on words matching the base characters before the '*' and to ignore any trailing characters. For example, you might use right truncation in a question such as "geo*", which may retrieve documents containing the words: geographer, geography, geologist, geometry, geometrical, etc.

Query Reporter

A query report is a document created by the server that describes how a client's question is parsed by the server. When a client asks a question of a database, the server creates and returns a query report to the client. The query report is the last document in the relevance-ranked list of documents returned by the server. The headline of the query report is listed as 'Query Report for this Search', and its relevance score is 1. Since the query report is an actual document, it may be retrieved for viewing by the client.

The query report contains the following information:

- The database being questioned,
- The original question,
- The Boolean equivalent of the question in infix notation (This notation is a fully-parenthesized version of the question, showing the Boolean operator precedence),
- The Boolean equivalent of the question displayed as a tree,
- The number of documents and the number of words in the database,
- The number of unique words in the database,
- The number of times each word in the question occurred in the database,
- The expanded search words resulting from right truncation, and
- The number of documents found that satisfied the question, and the amount of elapsed time it took to perform the search.

The purpose of this information is to give the user feedback on how the question was interpreted by the server, and on how well the information in the database matched the words in the question. Below is an example query report generated from the following question using right truncation and the AND Boolean operator: "carbon monox* AND poison". Simply stated, the question is looking for documents on carbon monox* and poison*. The question uses right truncation for monox* and poison* to match words such as monoxide, monoximes, etc., and poisoned, poisoning, and poisonous.

```
Headline: Query Report for this Search

This is the search report for the search you ran on Jan 11 13:31:51 1995.

It is a temporary file, and will expire about an hour after the search.

-----------------------------------------

Searching /wais/indexes/mydatabase...

Your query:

  carbon monox* AND poison*

is equivalent to:

  ((carbon monox*) AND (poison*))
```

```
and was interpreted as:
  AND
   ( carbon monox*
     poison*
   )
The database contains 39,062,401 words in 230,750 documents.
There are 639,200 different words.
carbon occurs 30,404 times in 14,896 documents.
monox* is expanded to:
  monoxide occurs 3,825 times in 2,515 documents.
  monoximes occurs 1 time in 1 document.
  monoxodithioacetal occurs 1 time in 1 document.
  monoxygenases occurs 2 times in 2 documents.
  monoxyhemoglobin occurs 1 time in 1 document.
poison* is expanded to:
  poisoned occurs 61 times in 49 documents.
  poisoning occurs 486 times in 283 documents.
  poisonous occurs 17 times in 14 documents.
The search found 67 documents.  It took about 5 seconds.
----------------------------------------
The search was performed by a WAIS Inc.server:  WAIS waisserver 1.0.
For more information email info@wais.com.
```

Security

Access List Security

The WAIS server uses an access-list security system to limit client access to WAIS databases. Before processing a client's request, the server checks to see if the requesting client has access to the requested database. It does this by checking an access list maintained for each database. The access list tells the server the legal client machines that have access to the database. The identity of the machine is based on the machine's Internet address. The access-list security system can be used with all existing WAIS clients, and is built in to the WAISserver software.

Authentication Option for the WAIS System

Kerberos is an authentication system designed by MIT for use on the Internet. MCC, the Austin-based consortium, is offering this as an add-on to WAIS Inc. products for those who would like this level of user authentication. Kerberos requires special client software, but uses WAIS Inc.'s server technology. Kerberos offers centralized key management for database holdings.

Encryption Option for the WAIS System

Public key authentication and encryption systems work well in the WAIS system. Encryption systems can be added to the WAIS system since the directory and architecture are compatible with the Whitfield Diffie public key system structure. With this structure, a variety of communications systems can be used without changing the encryption scheme. This facility will be built into the WAIS system based on customer demand and requires specialized client software.

Monitoring and Usage Reports

The WAIS server automatically records all transactions in a log file. The usage characteristics of your server can be extracted from this log file and summarized in a WAIS Usage Report.

For each client process requesting service, the WAIS server records in the log file the server's process identification number, the current count of the number of transactions performed for this client, the date, the time, and the type of transaction. The WAIS server records six transaction types:

- Opening a connection,
- Searching a database,
- Returning results from a search,
- Retrieving a document,
- Closing a connection, and
- Errors and warnings.

If a search transaction is performed, the server records the name of the database and the client's question. If results were returned from a search, the number of documents found and the document identifiers are logged. And finally, if a document is retrieved, the document identification number, the database name, the document size, and the document display format are all recorded.

A WAIS Usage Report summarizes the information contained in the log file. The summary contains the following information:

Total Number Of Connections

The total number of independent client connections made to the WAIS server. A single connection can span over multiple searches and retrievals, and over multiple databases.

Number of Different Machines Connecting

The total number of different client machines requesting services from this WAIS server.

Total Number of Searches

The total number of searches requested by all clients.

Total Connect Time (seconds)

The sum of the connect time of all clients. The connect time is the lifetime of each server process, in seconds. The majority of the connect time is idle time.

Total Search Time (seconds)

The sum of the search time of each server process, where the search time is the elapsed time, in seconds, that each server spends servicing its client's search request.

Searches Returning Zero Hits

The total number of search requests resulting in no matches, where the server process was unable to find any documents matching the client's question.

Total Number of Documents Retrieved

The total number of documents retrieved by all clients.

Total Number of Databases Searched

The total number of different WAIS databases that clients have searched.

Number of Searches with no Database Name

The total number of times clients requested a search without specifying the database name.

Number of Searches Requesting Help

The total number of times a client process requested a search for "?" or "help". This gives you an idea of how many new users are requesting information about the databases served on this machine.

Average Number of Seed Words per Search

The sum of the number of words contained in all questions divided by the number of questions, or search requests. The word count also includes Boolean operators and stopwords.

Number of Searches using Relevance Feedback

The total number of searches performed with relevance feedback.

Number of Server Warnings

The number of times a warning occurred while processing a client's request.

Number of Server Errors

The number of times an error occurred while processing a client's request.

Number of Search and Retrieval Requests

The total number of search and retrieval requests for each database searched by a client process. This information gives you a quantitative idea of the load on each database provided by the WAIS server.

List of Client Machines

The names of all client machines accessing the server's databases and the number of connections requested by each machine.

Client Information

The names of the client software and the number of connections requested by clients using this software.

Errors and Warnings

The error and warning messages of any problems reported by the server.

4

The WAIS Protocol Suite

WAIS and Standards

The WAIS protocol suite is based on national and international standards. WAIS Inc. is committed to using standards for communicating between clients and servers since this offers both technical and market-growth advantages. As different vendors offer WAIS-compatible systems for serving different markets, the protocol becomes the important piece for tying the implementations into a working whole. The WAIS protocol enables the interoperability of a global system made up of thousands of interacting pieces.

Because the WAIS protocol is based on open standards, more value can be brought to WAIS-compatible products from many companies. With companies in different fields all joining in to using the protocol, the expertise of each field is leveraged. For example, while software vendors develop and market user interfaces, publishers craft appropriate collections and presentations of information.

Several different standards all working together are needed to make a successful network publishing system since one standard would not take advantage of successful work being done on other standards. For example, WAIS is not restricted to using only one document format, since many existing document formats are in current use and are likely to change. Support for multiple standards also lessens reliance on any one proprietary standard. This prevents customers from being locked into a single vendor's proprietary system.

The WAIS protocol has been proven globally by WAIS Internet servers in 12 countries supplying over 900 databases, serving over 120,000 users in 28 countries. Given the current success of the WAIS protocol on the Internet, WAIS Inc. continues to support the standards and push for the enhancements needed for the network publishing industry.

Characteristics of the Protocol Suite

The characteristics of the WAIS protocol suite include:

Based on Open Standards

WAIS Inc. is committed to using national and international standards. The alternative to basing the system on open standards is to use proprietary standards. Where proprietary protocols may be more responsive to user needs because the authority to change them is centralized, the proprietary nature places reliance on a single company. Network publishing depends on a "critical mass" of publishers and users. Avoiding impediments in the way of creating that critical mass is essential to the long-term success of the system.

Scalable to Large Numbers of Distributed Clients and Servers

Internal to an organization, or across a wide-area network, the WAIS protocol is scalable to large numbers of distributed clients and servers. Finding the right server to be asking questions of is a large part of the challenge of a successful network publishing system. Users are informed about what servers are available, how to contact them, and how much they cost.

Scalable to Small and Large Data Collections

The size of a data collection can range from less than a megabyte to tens of gigabytes. A network publishing protocol allows searching and browsing in these environments. Hierarchical file browsing such as directories, FTP, or Gopher, do not scale to very large collections of data. Keyword searching does not extend to very small collections. A successful network publishing system incorporates both needs.

Provides Accounting Information for Billing

The exact billing method used in network publishing is left up to the network publisher. Pay-per-minute has been the traditional method on dialup systems, while subscriptions have been the method on most CD ROM releases. Pay-per-search or pay-per-retrieval have also been tried. With a client-server system, any combination can be used since the server monitors user activity.

Provides Access Restriction for Security and Charging

Any user might be able to get information about a database but be restricted from searching or retrieving information from it, or in other cases no information should be known to outside users. Varying levels of security is available on a database-by-database basis to address the concerns of security and charging.

Flexible to Adapt to the Changing Needs of a New Industry

The WAIS protocol suite is flexible enough to accommodate the needs of a rapidly changing industry. For example, it takes into account what is known about future needs such as video, wireless networks, and penpoint systems. This means that the standards communities behind the different pieces are responsive, and there is a mechanism for introducing new standards into the suite. WAIS Inc. works with both the network publishers and the standards committees to ensure an enduring environment for publishing.

Components of the Protocol Suite

The WAIS protocol suite is made up of several pieces:

Information Retrieval Standard Z39.50

The WAIS protocol is based on the ANSI (American National Standards Institute) NISO (National Information Standards Organization) Z39.50 Information Retrieval Service and Protocol Standard. The national standard is a superset of the ISO (International Standards Organization) Search and Retrieve Service Definition and Protocol Specification standard (ISO 10162 and 10163). Both the 1988 Version 1 and the 1992 Version 2 of the Z39.50 protocol are supported.

Document Formats

WAIS is used to find and retrieve information in many standard document formats. For example, word processor documents can be integrated into the same system with DBMS records and news feeds. Example formats are: SGML, GIF, MARC, ASCII, Microsoft Word, Microsoft Excel spreadsheets, and CAD drawings.

Document Identifiers

WAIS returns a list of document identifiers (doc-ids) as the result of a search. A doc-id describes how to retrieve a document's data. This allows a user to disconnect from the server and still be able to retrieve the document at some later time. The doc-id structure also allows a document to refer to another without having to republish it. This eliminates copyright violations by always pointing back to the original source and allowing the original owner to control access to the document. WAIS doc-ids were developed as part of the WAIS system because, at the time, there was no standard. WAIS Inc. is currently working with the IETF (Internet Engineering Task Force) on a new standard called the Universal Resource Locator and Universal Resource Name.

Server Descriptions for the Directory of Servers

To support a distributed system of WAIS servers and databases, the Directory of Servers was designed as a means of describing servers and making their databases publicly available. The Directory of Servers facility became a standardized part of the WAIS system. The Z39.50 committee is in the process of drafting a new Explain facility for servers. WAIS Inc. plans to incorporate the Explain facility as soon as it becomes a stable part of the Z39.50 standard.

New pieces that are being added include:

Authentication and Encryption

Authentication and encryption facilities are needed in highly secure environments. Password systems, unfortunately do not scale well into a client-server environment since a user would have to remember a large number of passwords, one for each server. The Kerberos system from MIT and public-key encryption systems offer a more scalable solution for the heterogeneous network environments available today. WAIS Inc. will be working on this approach.

Billing Format Standards

The server should feed usage information into billing systems. As standards appear in this environment WAIS Inc. will pursue them. At this time, each billing system requires a different report format, and custom WAIS tools.

Query Formats for Spatial Searching

The US Geological Survey is pursuing standards for searching areas of the globe using latitude and longitude and returning maps. There will be other special cases for searching that will be put on top of WAIS such as DNA searching, and these will require query formats as conventions and later as standards.

Hardcopy Delivery of Documents

Without widespread page image standards for computers, there is a growing demand for FAX and other hardcopy delivery facilities. For example, the mechanism by which a client asks for a document to be delivered via FAX is now being actively discussed.

Submitting Documents to a WAIS Database

The number of different types of databases that would work well with WAIS would be greatly increased if it were easy to submit documents. Standards for the submission of documents to WAIS servers are being addressed.

5

The Extendable WAIS Server

WAIS Inc. offers a variety of tools to broaden the reach of your WAISserver software. Here we describe some of WAIS Inc.'s most versatile additions to WAISserver software capabilities: WAISgate, Multi-databases, Filters, and Toolkits.

WAISgate Connects WAIS with Web

The World-Wide Web (WWW) is a popular Internet browsing and retrieval tool based on hypertext links between information sources. The WAIS Inc. WAISgate product is a gateway between WAIS and WWW. If you are serving a WAIS database, WAISgate can greatly expand your user base by making your WAIS database(s) accessible from the WWW. Similarly, if you are serving a set of WWW pages, WAISgate brings database searching capabilities to WWW client users.

Like WAIS, The World-Wide Web operates on a client-server model, but the two have different communication protocol and data format standards. WAISgate overcomes these differences. WAISgate receives WWW requests and forwards them to your WAIS server, effectively creating a gateway between a WWW Server and a WAIS server. WAISgate can handle requests from most WWW clients that use Forms. This includes X-Mosaic, the most widely used WWW client, and the character-based LYNX.

Serving Multi-Databases

The WAISserver software is capable of serving multiple databases as if they were one big database. A database configured in this manner is called a multi-database.

The benefits of a multi-database are numerous:

- **size and speed**

  A very large set of data is often more easily managed if split up between two or more directories — possibly on separate disks. In some cases, this becomes mandatory because the WAIS index files (especially **`.inv`** and **`.dic`**) approach or overflow the file limit for the UNIX implementation you are using. In other cases, splitting a database into two or more portions is desirable because it cuts down on indexing time and space requirements when updating or modification is required.

- **flexibility**

  To WAIS client users a multi-database appears to be a single database. However, depending on how it is set up, a multi-database can offer users the greater flexibility of searching each component database separately or searching them all as a multi-database. The WAIS administrator can choose whether or not to publicize the component databases by creating and entering their **`.src`** files in a directory of servers. Unless you inform them, users will assume the multi-database is simply another WAIS database.

- **accuracy**

  The WAISserver multi-database mechanism ensures that document relevance scoring is merged across the component database searches. Although a client process can search multiple databases without needing to have them defined as multi-databases, accurate score merging is not done by WAIS clients.

A multi-database is easily set up simply by creating a directory for the multi-database and putting two files in it: 1) a WAIS configuration file and, 2) a source file (`.src`).

Filters Extend Your WAIS Server

An external filter is a simple way to extend the functionality of your WAIS server without requiring source-code modifications. An external filter can be used to:

- Access data stored in a remote database (e.g., access data in an external DBMS),
- Change the display format of the document (e.g., convert TIFF to GIF or convert SGML to ASCII text),
- Allow the server to cache documents that are hard to generate (e.g., cache a document from an external database, or a document converted to a different display format)
- Restrict access of a document to only specific users (e.g., by modifying the document that the user receives),
- Create documents based on dynamic data (e.g., current temperature), and
- Provide out-of-band retrieval (e.g., FAX delivery of a document to the user).

A filter is a program run by the WAIS server when a WAIS client requests retrieval or relevance feedback of a document. The server invokes the filter with a document identifier (doc-id). The filter uses the doc-id to access and optionally modify the requested document, and sends the result back to the WAIS server. The server then either returns the result to the client, or uses it to perform relevance feedback.

Servers can be configured to run several filters. In addition, each database can define its own filters. The server checks a configuration file at access time to determine which filter to use. Since filters are external to the server, they can be customized and maintained by the system administrator. An external filter can be implemented in any standard UNIX programming language or shell script.

Custom Parser Toolkit

A large number of built-in parse formats are included in the WAISserver software. These parse formats can handle many standard file-based data formats. When a new file format is encountered, one of two strategies may be used to parse the new format. First, the data may be converted to one of the existing WAIS parse formats. Second, when conversion is not possible, a new parser can be developed to handle the new format. The Custom Parser Toolkit is a well documented C program that includes source code for existing parse formats along with instructions and pseudo code for any customized functions that may be needed in a new parse format.

Alternatively, WAIS Inc. can provide custom programming services to develop and maintain specialized parsers to meet the needs of individual publishers on a contract basis.

Client Toolkit

The Client Toolkit is designed to help client writers get a fast start when developing new WAIS clients. It contains source code and on-line documentation of an Application Programmers Interface (API) for running both Version 1 and Version 2 of the Z39.50 information retrieval protocols. Also included in the toolkit is a demonstration program that shows the use of the client protocol toolkit API.

WAIS Inc. urges client writers to update their clients to the new protocol version and to develop new clients using Z39.50V2. This new protocol standard includes the ability to retrieve an unlimited number of documents, whereas the older version has strict limits based on the protocol buffer size — generally around 200 documents.

6

The WAIS Forwarder

The WAIS Forwarder product, in conjunction with a "firewall" machine, provides access to external WAIS servers from within secure environments. The WAIS Forwarder is appropriate for secure sites connected to an external network, such as the Internet, through a firewall machine.

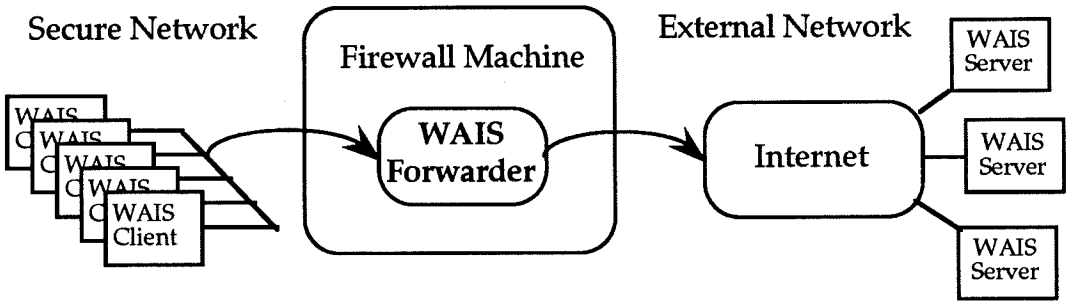

Figure 4: Configuration of the WAIS Forwarder

As shown in Figure 4, a firewall is a machine that connects a secure network to an external network. Information from one network destined for the other network must pass through the firewall machine. A forwarder is a software program running on the firewall machine that permits two application programs executing on either side of the firewall to communicate with each other. The forwarder allows machines on the secure network to access the services available on machines in the external network.

In a client-server application such as WAIS, a client contacts the forwarder on the firewall machine and the forwarder contacts the outside servers. Secure machines can open connections to Internet servers transparently by sending the request to the forwarder which automatically passes the request onto the external service. External machines cannot open connections to the forwarder, thus forming a one-way security system.

The WAIS Forwarder provides a secure network with all the benefits of the Internet WAIS servers without opening the secure network to external traffic. All WAIS functions are supported through the forwarder including the Directory of Servers, searching, and retrieval of text, images, and other formats. Because the WAIS Forwarder also forwards the IP address of the requesting client machine, databases using WAIS Inc. servers will continue to provide access-list security. In addition, the WAIS Forwarder optionally logs transaction statistics, enabling the firewall maintainer to monitor usage patterns.

The WAIS Forwarder is a software-only product that runs on many popular UNIX platforms and is easily configured and administered. In addition, the WAIS Forwarder works with all existing client software. For those that have special needs or security considerations, the product is available in source code as well as in executable form. The WAIS Forwarder can be purchased separately, or bundled with the WAIS Inc. server products. As new versions of the WAIS protocol suite come into widespread use, the package will be upgraded according to the maintenance and support agreement selected.

Appendix A
Glossary of WAIS Terms

.acc

File extension for access files. An access file contains the IP addresses of all machines that are allowed to search the database. It is created by the database administrator to control access to a database.

.cat

File extension for catalog files. See also **catalog**.

.dct

File extension for dictionary files. A dictionary file contains the dictionary of all the words used in a database.

.doc

File extension for document table files. A document table file contains a record of each document in the database.

.fn

File extension for filename table files. A filename table file lists the filenames of the original data files.

.hl

File extension for headline table files. A headline table file contains the headlines of all the documents in the database.

.inv

File extension for inverted files. An inverted file lists of all the words in the database, and for each word all the documents which contain that word.

.qst

File extension used by question files. See also **question structure**.

.src

File extension used by source description files which describe the database and the server. They are the means by which a client contacts the server and searches the

database. Source description files are distributed to clients by the directory-of-servers. They are ASCII files, and can be edited by hand. See also **source structure.**

Boolean operator

A Boolean operator is a mathematical operator based on set theory. Boolean operators provide a very powerful mechanism for specifying exact relationships between words, literal phrases, and field specifications in a WAIS query. Each Boolean operator denotes a specific relationship between the operands on either side of it. For example, the question **dog AND cat** contains the Boolean operator **AND**, and specifies that only documents having both the word **dog** and the word **cat** will be retrieved. In terms of set theory, this question is the intersection between the sets defined by **dog** and the sets defined by **cat**. The WAIS Server supports the following Boolean operators: AND, && (meaning AND), OR, || (meaning OR), ADJ (meaning adjacent to), and NOT.

catalog

Catalog (.cat) files contain a human readable list of headlines and document id's for some or all of the documents in the database. This list may be returned to a user whose search has gone poorly, as an aid to help them understand the contents of the database.

client

In a client/server architecture, the client is the program which requests services. With WAIS, clients are user interface programs which request services from remote or local services, using the WAIS protocol.

custom parser

A parse format created specifically to recognize a non-standard data format. A WAIS Server Administrator can write a custom parser with the aid of the Custom Parser Toolkit, provided with the WAISserver software. For a fee, this programming task can also be performed by WAIS Inc.

custom filter

See filter program.

daemon mode

This is the name for the mode the server is in when it is run by the **inetd** network daemon. See also standalone mode.

database

A service on a server that answers questions based on some defined expertise. It is described by a .src file that can usually be retrieved by asking "help" of the server now.

directory of servers

A database that serves .src files of WAIS databases. The public Directory of Servers on the Internet is located at wais.com on port 210.

display format

The display format specifies how a client will display a retrieved document. By default, the display format is determined by the parse format. See also **parse format.**

document identifier

(aka doc-id) A unique string that identifies each document in a database. A document identifier is assigned by the **waisparse** program. It includes the document key and resides in the filename table.

document key

(aka doc-key) A unique string used to locate each document in a database. If a database is file-based, then the document key is the full path and filename of the file containing a particular document. If a custom parser is used, it assigns document keys and these may be customized to suit your site's needs.

external database

Any non-WAIS database, such as an RDBMS, which has its own way of storing and retrieving data.

field

For data collections whose documents are structured in a semi-regular format, the regular portions of the documents can be tagged by the WAIS parser as fields. For example, in an electronic mail message, the 'to' and 'from' words are fields.

fielded search

Performing a restricted search based on the value of field or set of fields is called fielded search.

filter program

A custom software component that retrieves and modifies data as an enhancement to the waisserver program. Filter programs are most often used to convert display formats and to process data from external databases.

gwais

gwais is a WAIS freeware client program for GNU-Emacs developed by Jonathan Goldman at Thinking Machines (jonnyg@synopsys.com).

help

If a WAIS client user types **help** as a query, or issues an empty query, one or more text files are displayed. If the file **index.hlp** exists in the WAIS database index directory, it is displayed. Otherwise, the **.src** file is displayed (unless the System Administrator has invoked **waisindex** with the **-nocat** option). In addition to one or the other of these, the **.cat** file is also displayed(unless the System Administrator has invoked **waisindex** with the **-nosrc** option).

inetd daemon

The **inetd** daemon is the Internet network daemon. This is a UNIX program that manages the network services according to the configuration specified by the **/etc/inetd.conf** file. The **inetd** daemon runs in the background and manages network requests by spawning off other daemons, or programs, to service these requests.

multi-database

A multi-database is a WAIS database composed of two or more subsidiary WAIS databases. It is easily defined by making an entry in the WAIS configuration file, **index.wc**.

parse format

A parse format determines how **waisparse** breaks a data file into its component documents, and decides which words to use as a headline. See also **display format.**

phrase matching

If the user's question contains a phrase, i.e. a natural language expression made up of more than one word, a phrase-matching technique is employed. Using this technique, a higher weight is assigned to a document containing a phrase that identically matches the phrase in the user's question.

plural stemming

Plural stemming is a stemming algorithm that attempts to derive the root form of a word if given a plural.

port

A port number specifies a service on a UNIX machine. For example, WAIS usually runs on port 210, and Telnet usually runs on port 23. Ports below 1,000 are called "well-known" ports, and you must be superuser to run a program on them. The mapping between the service name and port number is determined in /etc/services file.

porter stemming

Porter stemming is a stemming algorithm that attempts to find the real base, or stem, of a word and derive any possible alternate variations.

proximity relationship

A proximity relationship designates that if the words in a question are located close together in a document, they are given a higher weight than those found further apart. The idea behind a proximity relationship is that words found in close proximity to each other in a document more likely contain the same content as that specified in the user's question.

query

See **question.**

query report

A query report is a document created by the server that describes how a client's question is parsed by the server.

question structure

A question structure is a file format used by some client programs to store the state of a user's inquiry so that it can be used again. It includes what databases to ask, the user's question, and any relevant documents. It may also store the last list of results.

question

A question is an expression containing a combination of natural language words and Boolean operators.

relevance feedback

Relevance feedback is the ability to select a document or a portion of a document and find a set of documents similar to the selection. In essence, relevance feedback adds more words to the original question. It uses the most important words and phrases in the relevant document in addition to the original question. The most important words and phrases are determined by the same weighting algorithms as the words in the original question. The weight of the relevant document terms is less than the original question terms.

relevance ranking

Relevance ranking is the scoring of documents based on their relevance to a user's question, where the most relevant document has the highest score, or rank. A document receives a higher score if the words in the question are in the headline, or if the words appear many times, or if the phrases occur exactly as in the question. A document's score is derived from using techniques such as word weighting, term weighting, phrase matching, proximity relationships, and word density.

results list

The results list is the first thing returned to a client user in response to a query. It contains a list of headlines for the documents found along with the relevance rank or "score" assigned to each document. The client user may request retrieval of the listed documents, one at a time.

right truncation

A user can specify right truncation in a question by ending a word on the right with the wildcard character, '*'. This tells the WAIS server to search on words matching the base characters before the '*'. An example of using right truncation is a question such as "geo*", which may retrieve documents containing the words: geographer, geography, geologist, geometry, geometrical, etc.

serve

The transitive verb "to serve" is used when describing the services provided by a WAIS server. The server serves data to the client.

server

In a client/server architecture, the server is the program that provides the services requested by the client program. The machine on which the WAIS server software is installed is also know as the server.

Solaris

(aka. SunOS 5.x) Sun's version of AT&T's System V Release 4 UNIX operating system.

source structure

A source structure is a file format used by WAIS client programs to describe a particular database on a particular server. Typical contents are the machine name, IP address, port of the server, and the name of the database.

standalone mode

A WAIS server is run in standalone mode when the **waisserver** program is invoked directly by the user or shell script. See also daemon mode.

stemming

Stemming is a technique used to automatically derive the root of a queried word. The root is then used to search against the roots of the words contained in a database. If a question contains the word "skate", for example, stemming is used to find documents that may also include "skated" and "skating".

stopword

A stopword is a word that occurs so frequently that it is not useful for distinguishing one document from another. Since it is not useful for searching, it is not indexed.

superuser

On machines running the UNIX operating system, the superuser is a privileged user who has access to all files and all services provided by the machine. This status is usually reserved for system administrators.

swais

swais is a freeware WAIS client program for ASCII terminals developed by John Curran.

TCP/IP

TCP/IP is an acronym for Transmission Control Protocol/Internet Protocol. This is the low level protocol which is used on the Internet, and on many LANs. It provides reliable data communication.

term weight

Each word used in a database is assigned a numerical value, called the term weight, based on the frequency of occurrence of that word over all documents in the database. Words that occur frequently throughout the database are not weighted as highly as the

words that appear less frequently. Very common words are either ignored or diminished in the scoring. For example, since the term, "animal", may occur frequently in many of the documents in a database, its term weighting is small compared to a term such as "hippopotamus", which may occur in only a small number of documents.

WAIS

WAIS is an acronym for Wide Area Information Servers and a trademark of WAIS Inc.

WAIS database

A WAIS database consists of a set of documents, a set of index files, and a source file. See also **database.**

WAISgate

The WAISgate product is a gateway between WAIS and the World-Wide Web, an Internet browsing and retrieval tool based on hypertext links between information sources.

WAIS protocol

The WAIS protocol is used to connect WAIS clients and servers. It is based on the Z39.50 protocol. Because a standard protocol is used, clients and servers can be built on a wide variety of computer architectures communicating over local and wide-area networks.

waisindex

This program takes the documents specified by waisparse and builds a WAIS searchable index.

waislookup

This program is essentially a very simple WAIS client generally used for testing a WAIS installation. It provides an interactive, dumb terminal interface. It can be invoked as an interactive server, in which case it does not go through the protocol or server but rather does the search internally. This mode is useful for troubleshooting WAIS indexes. The waislookup program can also be invoked as a client and then communicate with either a local or remote WAIS server. This mode is useful for troubleshooting WAIS servers.

waisparse

This program reads a data file and breaks it into its component documents, decides which words to index, and which words to use as the headline. The output of waisparse is fed to waisindex.

waisreporter

This program summarizes the log files generated by waisserver. It can be used to monitor database usage.

waisserver

This program waits for a connection from a client, and handles the connection by searching and/or retrieving documents from a WAIS database.

word density

The ratio of the number of times a queried word appears in a document to the size of the document is called the word density. It is a measure of how important a queried word is to the overall content of the document. A higher word density results in a higher relevance ranking.

word weight

If a word in a document is found to match a word in the user's question, the word is assigned a word weight, and this weight additively contributes to the overall score of the document. The exact weight that a word receives depends on where in the document the word was found. A word is weighted highest if it appears in the headline, less if the word appears in all capital letters or if the first letter of the word is capitalized, and finally, a word has the least weight if it appears only in the text.

xwais

xwais is a freeware WAIS client program for X-Windows developed by Jonathan Goldman at Thinking Machines (jonnyg@synopsys.com).

Z39.50

Z39.50 is the National Information Standards Organization (NISO) protocol for information search and retrieval.

Appendix B
WAIS References

Hard copies of most of the following documents are available from WAIS Inc. Some documents are available electronically via anonymous ftp or gopher at ftp.wais.com, but may not contain figures in the ASCII version. Email, FAX, mail, or phone your name, address, email and phone number to: WAIS Inc, 1040 Noel Drive, Menlo Park, CA, 94025, phone: 415-617-0444, FAX: 415-327-6513, email: info@wais.com

WAIS Articles and Publications

Britannica's 44 Million Words are Going On Line, New York Times, John Markoff, February 8, 1994, pp C1.

Curtain's Rising on a Third Generation of On-Line Services, New York Times, John Markoff, January 30, 1994.

Pointing Finger, WAIS at Internet Addresses, MacWeek, Jeff Ubois, May 28, 1993, pp 42, 44.

WAIS Offers Publishing Products, Open Systems Today, Paul Kapustka, May 10, 1993, pp 13.

Unix Servers Distribute On-line Information, Info World, Cheryl Gerber, May 3, 1993, pp 6.

Info Access Plan Promises Power to Fed Users, Federal Computer Week, Jennifer Jones, March 29, 1993, pp 1, 41.

A Web of Networks, an Abundance of Services, New York Times, John Markoff, February 28, 1993.

Good-bye, Dewey Decimals, Forbes Magazine, David Churbuck, February 15, 1993, pp 204-205.

Internet Retrieval Tools Go on Market, Network World, Ellen Messmer, February 15, 1993, pp 29, 77.

Federal Information on the Internet, Anna Keller, Library of Congress, February 1993.

Internet of the Future may be a One-Stop Information Shop, MacWeek, Margie Wylie, January 25, 1993, pp 22, 24.

Index Everything, Share It Companywide with WAIS, MacWeek, Daniel P. Dern, October 26, 1992, pp 24-25.

Help is on the WAIS, American Libraries, M. Lukanuski, October 1992, pp 742-744. Feature article with some pros and cons of the WAIS protocol from the library community point of view.

Information - the Commodity of the Future, Merit/NSFNET Link Letter Newsletter, Merit/NSFNET Information Services, September/October 1992. Follow-up to above article, explaining how Merit/NSFNET is utilizing the different information services available. Available via anonymous ftp: /pub/wais-doc/linkletter2@ftp.wais.com.

Identifying and Describing Federal Information Inventory/Locator Systems: Design for Networked-Based Locators, Charles R. McClure, Joe Ryan, and William E. Moen, School of Information Studies, Syracuse University, August 25, 1992, volume 1.

A Comparison of Internet Resource Discovery Approaches, M. Schwartz, A. Emtage, B. Kahle, B.C. Neuman, August 1992. Paper to appear in Computing Systems 5(4), 1992. In-Depth overview and comparison of current Internet information systems. Postscript copy available via anonymous ftp: /pub/wais-doc/resource-compar@ftp.wais.com.

WAIS: The Wide Area Information Server or Anonymous What???, Peter Marshall, June 18, 1992. Describes and details the implementation of WAIS at the University of Western Ontario. Available via anonymous ftp: /pub/wais-doc/UWO-wais-paper.ps@ftp.wais.com.

Personal Computing: Collective Dynabases, Communications of the ACM, Larry Press, June 1992, pp 26-32. Overview of WAIS and commercial projects.

WAIS: A New Development in Information Services, MIT I/S N Newsletter, T. MacRae and S. Jones, June 1992. Overview of WAIS by the Network Services and Publication Services at MIT. Available via anonymous ftp: /pub/wais-doc/MIT.IS.news@ftp.wais.com.

WAIS: Wide Area Information Servers, Information Intelligence Inc, George S. Machovec, March 1992, pp 1-5. Overview of WAIS from a librarian perspective. Available via anonymous ftp: /pub/wais-doc/lib.perspective@ftp.wais.com.

WAIS - Making it Easier to Access Internet Resources, Merit/NSFNET Link Letter Newsletter, Brewster Kahle, March/April 1992. Overview of WAIS. (reprinted from CERFnet News, Volume 3 Number 6) Available via anonymous ftp: /pub/wais-doc/linkletter@ftp.wais.com.

WAIS: Is it the Lotus 1-2-3 of the Internet?, Communications Week, Carl Malamud, March 16, 1992, pp 17. Brief article of WAIS on the Internet.

Where There's a Will, There's a WAIS, Digital Media - A Seybold Report, Denise Caruso, February 17, 1992, pp 5-6. Article touching on several issues of WAIS, such as protocol, client-server relationship, "for pay" servers, and legal issues.

The Reading Room, Digital Media - A Seybold Report, Brewster Kahle, February 17, 1992, pp 7-8. Essay on the controversy between private ownership of information and public access to information.

The Promise of the WAIS Protocol, UNIX Today!, Jason Levitt, December 9, 1991, pp 44, 47-48. Article describing the freeware release.

The Global Village Starts with WAIS, Tomaso Poggio, December 1991, Overview of WAIS in Italian.

Network to Unite Data Bases, San Jose Mercury News, John Markoff, July 21, 1991, pp 1F. Rewriting of the "For the PC User, Vast Libraries," New York Times article with emphasis on Apple component.

For Shakespeare, Just Log On, New York Times, John Markoff, July 3, 1991, pp C1. Overview of WAIS Internet experiment.

Browsing Through Terabytes, Byte Magazine, Richard Stein, May 1991, pp 157-164. Article on large WAIS systems.

WAIS Promises Easy Text Retrieval, MacWeek, Henry Norr, May 14, 1991, pp 22. Report on the Peat Marwick WAIS system.

Release 1.0, Release 1.0, Esther Dyson, April 1991, entire issue. In-depth article on commercial systems and protocols, featuring WAIS. (Hardcopy copies available from: EDventure Holdings, 375 Park Ave., New York, NY 10152; (212) 758-3434) Available via anonymous ftp: /pub/wais-doc/release1.0@ftp.wais.com.

Designing a Desktop Information System: Observations and Issues, Thomas Erickson & Gitta Salomon, Apple Computer. Human Factors in Computing Systems, CHI '91 Conference Proceedings, pp 49-54, April 1991, New Orleans. New York: ACM, 1991. Early paper on the Apple interface for WAIS.

An Analysis of the Effects of Data Corruption on Text Retrieval Performance, S. Smith, C. Stanfill, December 1988. Thinking Machines Corporation Technical Report TMC-68.

WAIS Videos

Special Interest Group on Wide Area Information Servers: Conference Held March 19, 1993, Open-File Report 93-252, USGS video on WAIS, VHS videotape $20. Send orders to Book and Open-File Report Sales, USGS, Federal Center, Box 25286, MS 306, Denver, Colorado, 80225.

Wide Area Information Servers Class: Indexer and Server, Open-File Report 93-253, USGS training video on WAIS, VHS videotape, 2-tape set $40. Send orders to Book and Open-File Report Sales, USGS, Federal Center, Box 25286, MS 306, Denver, Colorado, 80225.

Macintosh Demonstration Screen-Movie, Steve Cisler put together a short screen-recorder movie of WAIStation. Available via anonymous ftp: /pub/wais-doc/WAIStation-Canned-Demo.sit.hqx@ftp.wais.com.

Electronic Services

wais-discussion@wais.com: Biweekly digest of mail from users and developers on Electronic Publishing. If you have WAIS-related news please send electronic mail to wais-discussion@wais.com. Send requests for inclusion on the mailing list to wais-discussion-request@wais.com. Anonymous ftp access to archives: /pub/mail-archives/wais-discussion/issue-*@ftp.wais.com. Archives are available on the public WAIS database: wais-discussion-archives.src.

wais-talk@wais.com: An interactive list of developers that generates a couple messages a day. Send requests to wais-talk-request@wais.com. Archives are available on the public WAIS database: wais-talk-archives.src.

comp.infosystems.wais: A netnews discussion group on WAIS issues. All postings to wais-discussion@wais.com go to this group as well.

Z3950iw: Z39.50 implementors list for low-level discussions of protocol details. Send requests to listserv@nervm.nerdc.ufl.edu.

WAIS Freeware Information

Due to limited staff resources, WAIS Inc. cannot provide support for the freeware WAIS clients and servers. For information on WAIS server freeware or the Clearinghouse of Networked Information Discovery and Retrieval (CNIDR), contact Jane Smith at jane.smith@cnidr.org or 919-248-9213. The director of the freeware is George Brett at ghb@jazz.concert.net or 919-962-1000. For information on WAIS clients, contact the individual developers of each respective client.

WAIS Freeware Servers

Server	FTP Location
NeXT	/pub/freeware/next/*@ftp.wais.com
RS6000	/pub/freeware/rs6000/*@ftp.wais.com
SGI	/pub/freeware/sgi/*@ftp.wais.com
Source Code	/pub/freeware/unix-src/wais-8-*.tar.Z@ftp.wais.com
SUN	/pub/freeware/sun/*@ftp.wais.com

WAIS Freeware Clients

Client	Author	FTP Location
DOS	Jim Fullton, UNC	/pub/wais/DOS/*@sunsite.unc.edu, or /pub/tcpip/pcwais.zip@hilbert.wharton.upenn.edu
GWAIS (Gnu Emacs)	Jonathon Goldman Thinking Machines Corp.	/pub/freeware/unix-src/wais-8-*.tar.Z@ftp.wais.com
IBM Mainframe	Tim Gauslin, USGS	/pub/freeware/ibm-mvs/*@ftp.wais.com
Mac	MCC Francois Schiettecatte	/pub/freeware/mac/MacWAIS*@ftp.wais.com /pub/freeware/mac/WAISBrowser*@ftp.wais.com
Mac HyperCard	Francois Schiettecatte	/pub/freeware/mac/HyperWais*@ftp.wais.com /pub/freeware/mac/JFIFBrowser*@ftp.wais.com
Mail	Jonathon Goldman Thinking Machines Corp.	send message to waismail@quake.think.com, "search <source-name> {keywords}" or "retrieve DOCID" (DOCID as returned by a search)
NeXT	Paul Burchard, Univ of Utah	/pub/freeware/next/*@ftp.wais.com
Openlook	Simon Spero, UNC	/pub/freeware/open-look/*@ftp.wais.com
OS2	Kevin Oliveau, WAIS Inc Julie Mills, Library of Congress	/pub/freeware/os2/*@ftp.wais.com
SunView		/pub/wais/sunview/*@sunsite.unc.edu
SWAIS	John Curran, BBN	/pub/freeware/unix-src/wais-8-*.tar.Z@ftp.wais.com
Telnet Access	(uses SWAIS)	Telnet wais.com, login wais, password user@host
VMS	Jim Fullton, UNC	/pub/wais/vms/*@sunsite.unc.edu
Windows	Tim Gauslin, USGS MCC	/pub/freeware/windows/wnwais*.zip@ftp.wais.com /pub/freeware/windows/EIWAIS*@ftp.wais.com
XWAIS	Jonathan Goldman Thinking Machines Corp.	/pub/freeware/unix-src/wais-8-*.tar.Z@ftp.wais.com

Z39.50 Information and Publications

Z39.50-1988: Information Retrieval Service Definition and Protocol Specification for Library Applications. National Information Standards Organization (Z39), P.O. Box 1056, Bethesda, MD 20817. (301) 975-2814. Available from Document Center, Belmont, CA. Telephone 415-591-7600.

Z39.50-1992 (Version 2) ANSI Z39.50: Information Retrieval Service and Protocol, Final Text, July 1992. National Information Standards Organization (Z39), P.O. Box 1056, Bethesda, MD 20817. (301) 975-2814. Available from Transitions, 908-932-2280.

Z39.50-1991 Version 2, Draft 3, May 1991. Electronic version of the working copy of the Z39.50 implementors group. Available via anonymous ftp /pub/protocol/z3950-v2d3.txt@ftp.wais.com.

Z39.50-1992 Version 3, Draft 7, July 1993. Electronic version of the working copy of the Z39.50 implementors group. Available via anonymous ftp: /pub/protocol/z3950-v3d7.txt@ftp.wais.com.

The Z39.50 Information Retrieval Protocol: An Overview and Status Report, Clifford Lynch, Computer Communication Review ACM SIGCOMM.

The Z39.50 Protocol in Plain English, Clifford Lynch. Fall 1992. Available via anonymous ftp: /pub/protocol/plain.english@ftp.wais.com.

Internet Information

Internet Starter Kit for Macintosh, Adam Engst, Hayden Books, 1993. Contains freeware/shareware tools disk for Mac internet users.

The Mac Internet Tour Guide: Cruising the Internet the Easy Way, Michael Fraase, Ventana Press, 1993

The Whole Internet: User's Guide & Catalog, Ed Krol, O'Reilly & Associates Inc, 1992. (Chapter 12 entitled "Searching indexed databases: WAIS")

Exploring the Internet: A Technical Travelogue, Carl Malamud, Prentice Hall, 1992.

Internet Access Providers in the United States, The general types of services they provide, and how to contact them. From Chapter 4 of the book, "Internet: Getting Started". For more information about "Internet: Getting Started", contact SRI International at 415-859-3695, nisc@nisc.sri.com.

Internet Access Providers outside the United States. From Chapter 7 of the book, "Internet: Getting Started". For more information about "Internet: Getting Started", contact SRI International at 415-859-3695, nisc@nisc.sri.com.

Public Dialup Internet Access List (PDIAL), February 1994. A list of public access service providers offering dialup access to outgoing Internet connections such as ftp and telnet. Available by sending electronic mail to "info-deli-server@netcom.com", with the message subject "send PDIAL".